ABSTRACT

A system and method efficiently and anonymously retrieves large scale Web data through a restricted query interface. A number of proxy servers are utilized to permit parallel access to a target Web server for processing multiple queries simultaneously. Latency in the individual queries is absorbed by the proxy servers. Queries that would otherwise appear structured to the target server are assigned to the proxy server in a random fashion, obscuring the structured nature of the queries. The anonymous nature of the queries made by the proxy servers furthermore conceals the identity of the originating server.